Exploratory Data Analysis

Lecture 8

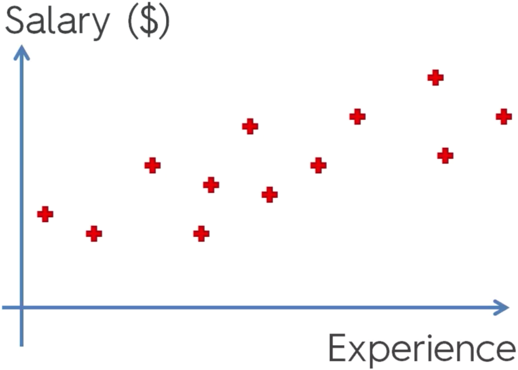
Corina Besliu

Technical University of Moldova

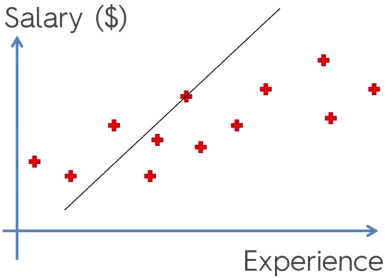
September 23, 2021

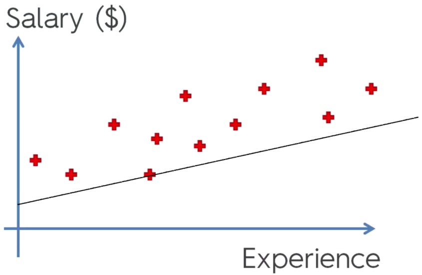
Salary and Work Experience

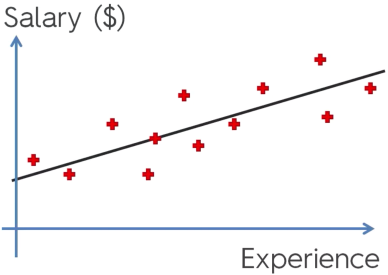
Fit a line to describe the relationship between Experience and Salary



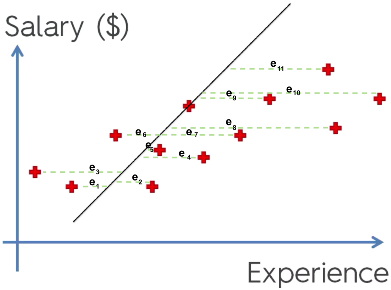
But what is the best line?

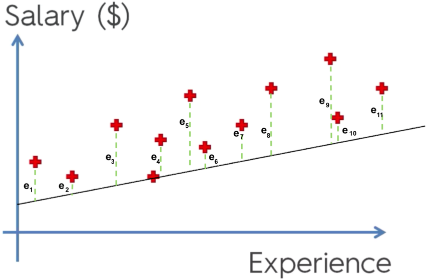


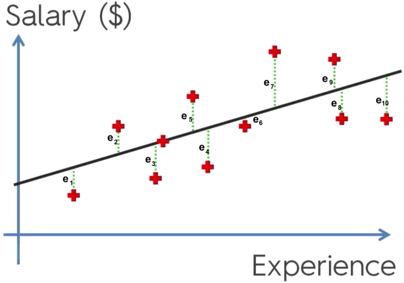




Choose the one that minimizes the residuals $\sum_{i=1}^{n} e_n^2$

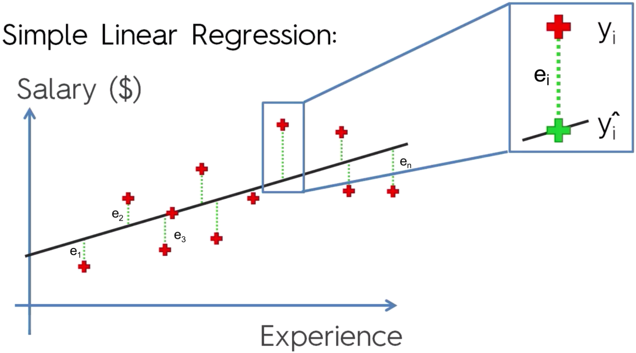






Quntifying the relationship

$$\min \rightarrow \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

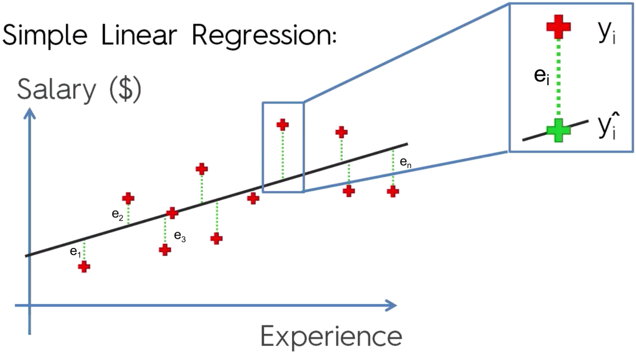


$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

$$\min \rightarrow \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

Quntifying the relationship

$$\min \rightarrow \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

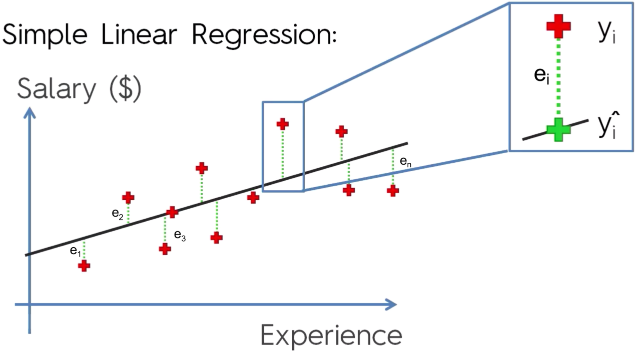


$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

$$\min \rightarrow \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

Quntifying the relationship

$$\min \rightarrow \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$



$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

$$\min \rightarrow \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

$\hat{\beta}_0$ and $\hat{\beta}_1$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} [(X_i - \overline{X})(Y_i - \overline{Y})]}{\sum_{i=1}^{N} (X_i - \overline{X})^2}$$

Salary and Work Experience

	Y	X
Individual	Salary (in thousand USD)	Experience (in years)
1	50	2
2	30	1
3	60	3
4	65	4
5	30	0

$\bar{Y} = 47$ $\bar{X} = 2$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}[(X_i - \bar{X})(Y_i - \bar{Y})]}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}[(X_i - \bar{X})(Y_i - \bar{Y})]/(n-1)}{\sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)} = \frac{Cov(X,Y)}{Var(X)} = 10$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 27$$

$$\widehat{Salary}_i = 27 + 10 * Experience_i$$

Salary and Work Experience

	Y	X
Individual	Salary (in thousand USD)	Experience (in years)
1	50	2
2	30	1
3	60	3
4	65	4
5	30	0
	$\bar{Y} = 47$	$\bar{X} = 2$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}[(X_i - \bar{X})(Y_i - \bar{Y})]}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}[(X_i - \bar{X})(Y_i - \bar{Y})]/(n-1)}{\sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)} = \frac{Cov(X,Y)}{Var(X)} = 10$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 27$$

$$\widehat{Salary}_i = 27 + 10 * Experience_i$$

Salary and Work Experience

Individual	Salary (in thousand USD)	Experience (in years)
	Y	X
1	50	2
2	30	1
3	60	3
4	65	4
5	30	0
	$\bar{Y} = 47$	$\bar{X} = 2$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}[(X_i - \bar{X})(Y_i - \bar{Y})]}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}[(X_i - \bar{X})(Y_i - \bar{Y})]/(n-1)}{\sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)} = \frac{Cov(X,Y)}{Var(X)} = 10$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 27$$

$$\widehat{Salary}_i = 27 + 10 * Experience_i$$

Salary and Work Experience

	Y	X
Individual	Salary (in thousand USD)	Experience (in years)
1	50	2
2	30	1
3	60	3
4	65	4
5	30	0
	$\bar{Y} = 47$	$\bar{X} = 2$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}[(X_i - \bar{X})(Y_i - \bar{Y})]}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}[(X_i - \bar{X})(Y_i - \bar{Y})]/(n-1)}{\sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)} = \frac{Cov(X,Y)}{Var(X)} = 10$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 27$$

$$\widehat{Salary}_i = 27 + 10 * Experience_i$$

Salary and Work Experience

Individual	Salary (in thousand USD) Y	Experience (in years) X
1	50	2
2	30	1
3	60	3
4	65	4
5	30	0
	$\bar{Y} = 47$	$\bar{X} = 2$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}[(X_i - \bar{X})(Y_i - \bar{Y})]}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}[(X_i - \bar{X})(Y_i - \bar{Y})]/(n-1)}{\sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)} = \frac{Cov(X,Y)}{Var(X)} = 10$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 27$$

$$\widehat{Salary}_i = 27 + 10 * Experience_i$$

Terminology

X - Independent Variable, Regressor, Predictor

Y - Dependent Variable, Regressand, Predicted Variable

β_0 - Constant, Intercept

β_1 - Regression Coefficient, Slope Coefficient

Terminology

X - Independent Variable, Regressor, Predictor

Y - Dependent Variable, Regressand, Predicted Variable

β_0 - Constant, Intercept

β_1 - Regression Coefficient, Slope Coefficient

Terminology

X - Independent Variable, Regressor, Predictor

Y - Dependent Variable, Regressand, Predicted Variable

β_0 - Constant, Intercept

β_1 - Regression Coefficient, Slope Coefficient

Terminology

X - Independent Variable, Regressor, Predictor

Y - Dependent Variable, Regressand, Predicted Variable

β_0 - Constant, Intercept

β_1 - Regression Coefficient, Slope Coefficient

Multivariate OLS

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + \epsilon_i$$

where:

i, goes from 1 to N and indicates the observation number

X_{1i} - indicates the ith observation of independent variable X_1

X_{2i} indicates the ith observation of another independent variable, X_2.

- The biggest difference with multivariate regression model is in the interpretation of the slope coefficients.
- Coefficients are called partial regression coefficients.
- Allow a researcher distinguish the impact of one variable from that of other independent variables.

Multivariate OLS

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + \epsilon_i$$

where:

i, goes from 1 to N and indicates the observation number

X_{1i} - indicates the ith observation of independent variable X_1

X_{2i} indicates the ith observation of another independent variable, X_2.

- The biggest difference with multivariate regression model is in the interpretation of the slope coefficients.
- Coefficients are called partial regression coefficients.
- Allow a researcher distinguish the impact of one variable from that of other independent variables.

Multivariate OLS

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + \epsilon_i$$

where:

i, goes from 1 to N and indicates the observation number

X_{1i} - indicates the ith observation of independent variable X_1

X_{2i} indicates the ith observation of another independent variable, X_2.

- The biggest difference with multivariate regression model is in the interpretation of the slope coefficients.
- Coefficients are called partial regression coefficients.
- Allow a researcher distinguish the impact of one variable from that of other independent variables.

Multivariate OLS

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + \epsilon_i$$

where:

i, goes from 1 to N and indicates the observation number

X_{1i} - indicates the ith observation of independent variable X_1

X_{2i} indicates the ith observation of another independent variable, X_2.

- The biggest difference with multivariate regression model is in the interpretation of the slope coefficients.
- Coefficients are called partial regression coefficients.
- Allow a researcher distinguish the impact of one variable from that of other independent variables.

Multivariate OLS

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + \epsilon_i$$

where:

i, goes from 1 to N and indicates the observation number

X_{1i} - indicates the ith observation of independent variable X_1

X_{2i} indicates the ith observation of another independent variable, X_2.

- The biggest difference with multivariate regression model is in the interpretation of the slope coefficients.
- Coefficients are called partial regression coefficients.
- Allow a researcher distinguish the impact of one variable from that of other independent variables.

Multivariate OLS

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + \epsilon_i$$

where:

i, goes from 1 to N and indicates the observation number

X_{1i} - indicates the ith observation of independent variable X_1

X_{2i} indicates the ith observation of another independent variable, X_2.

- The biggest difference with multivariate regression model is in the interpretation of the slope coefficients.
- Coefficients are called partial regression coefficients.
- Allow a researcher distinguish the impact of one variable from that of other independent variables.

Multivariate OLS

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki} + \epsilon_i$$

where:

i, goes from 1 to N and indicates the observation number

X_{1i} - indicates the ith observation of independent variable X_1

X_{2i} indicates the ith observation of another independent variable, X_2.

- The biggest difference with multivariate regression model is in the interpretation of the slope coefficients.
- Coefficients are called partial regression coefficients.
- Allow a researcher distinguish the impact of one variable from that of other independent variables.

Multivariate OLS Coefficient Interpretation

Specifically, a **multivariate regression coefficient** indicates the change in the dependent variable associated with a one-unit increase in the independent variable in question, *holding constant the other independent variables in the equation.*

Thus:

- The coefficient $\beta 1$ measures the impact on Y of a one-unit increase in X_1, holding constant X_2, X_3, . . . and X_K
- but not holding constant any relevant variables that might have been omitted from the equation (e.g., X_{k+1}).

Multivariate OLS Coefficient Interpretation

Specifically, a **multivariate regression coefficient** indicates the change in the dependent variable associated with a one-unit increase in the independent variable in question, *holding constant the other independent variables in the equation.*

Thus:

- The coefficient $\beta 1$ measures the impact on Y of a one-unit increase in X_1, holding constant X_2, X_3, . . . and X_K
- but not holding constant any relevant variables that might have been omitted from the equation (e.g., X_{k+1}).

Multivariate OLS Coefficient Interpretation

Specifically, a **multivariate regression coefficient** indicates the change in the dependent variable associated with a one-unit increase in the independent variable in question, *holding constant the other independent variables in the equation.*

Thus:

- The coefficient $\beta 1$ measures the impact on Y of a one-unit increase in X_1, holding constant X_2, X_3, . . . and X_K
- but not holding constant any relevant variables that might have been omitted from the equation (e.g., X_{k+1}).

Multivariate OLS

$$\mathrm{FINAID}_i = \beta_0 + \overset{-}{\beta_1}\mathrm{PARENT}_i + \overset{+}{\beta_2}\mathrm{HSRANK}_i + \epsilon_i$$

where:

- $FINAID_i$ = the financial aid (measured in dollars of grant per year) awarded to the ith applicant
- $PARENT_i$ = the amount (in dollars per year) that the parents of the ith student are judged able to contribute to college expenses
- $HSRANK_i$ = the ith students GPA in high school, measured as a percentage (ranging from a low of 0 to a high of 100)

Multivariate OLS

$$\mathrm{FINAID}_i = \beta_0 + \overset{-}{\beta_1}\mathrm{PARENT}_i + \overset{+}{\beta_2}\mathrm{HSRANK}_i + \epsilon_i$$

where:

- $FINAID_i$ = the financial aid (measured in dollars of grant per year) awarded to the ith applicant
- $PARENT_i$ = the amount (in dollars per year) that the parents of the ith student are judged able to contribute to college expenses
- $HSRANK_i$ = the ith students GPA in high school, measured as a percentage (ranging from a low of 0 to a high of 100)

Multivariate OLS

$$\text{FINAID}_i = \beta_0 + \overset{-}{\beta_1}\text{PARENT}_i + \overset{+}{\beta_2}\text{HSRANK}_i + \epsilon_i$$

where:

- $FINAID_i$ = the financial aid (measured in dollars of grant per year) awarded to the ith applicant
- $PARENT_i$ = the amount (in dollars per year) that the parents of the ith student are judged able to contribute to college expenses
- $HSRANK_i$ = the ith students GPA in high school, measured as a percentage (ranging from a low of 0 to a high of 100)

Multivariate OLS

$$\mathrm{FINAID}_i = \beta_0 + \overset{-}{\beta_1}\mathrm{PARENT}_i + \overset{+}{\beta_2}\mathrm{HSRANK}_i + \epsilon_i$$

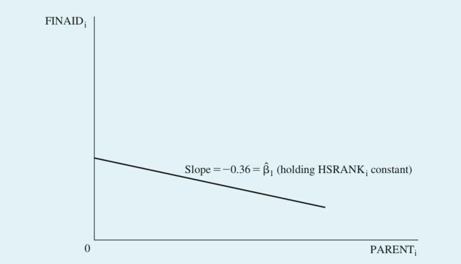
where:

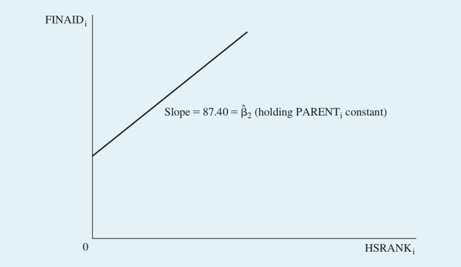
- $FINAID_i$ = the financial aid (measured in dollars of grant per year) awarded to the ith applicant
- $PARENT_i$ = the amount (in dollars per year) that the parents of the ith student are judged able to contribute to college expenses
- $HSRANK_i$ = the ith students GPA in high school, measured as a percentage (ranging from a low of 0 to a high of 100)

Multivariate OLS

- $FINAID_i$ = the financial aid (measured in dollars of grant per year) awarded to the ith applicant
- $PARENT_i$ = the amount (in dollars per year) that the parents of the ith student are judged able to contribute to college expenses
- $HSRANK_i$ = the ith students GPA in high school, measured as a percentage (ranging from a low of 0 to a high of 100)

$$\widehat{\text{FINAID}}_i = 8927 - 0.36\text{PARENT}_i + 87.4\text{HSRANK}_i$$





$\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$

$$\hat{\beta}_1 = \frac{\left(\sum yx_1\right)\left(\sum x_2^2\right) - \left(\sum yx_2\right)\left(\sum x_1x_2\right)}{\left(\sum x_1^2\right)\left(\sum x_2^2\right) - \left(\sum x_1x_2\right)^2}$$

$$\hat{\beta}_2 = \frac{\left(\sum yx_2\right)\left(\sum x_1^2\right) - \left(\sum yx_1\right)\left(\sum x_1x_2\right)}{\left(\sum x_1^2\right)\left(\sum x_2^2\right) - \left(\sum x_1x_2\right)^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}_1 - \hat{\beta}_2\bar{X}_2$$

How do we judge the goodness of our model?

Some concepts to help us judge how much of the variation of the dependent variable is explained by our regression.

$$\mathrm{TSS} = \sum_{i=1}^{N} (Y_i - \bar{Y})^2$$

TSS Decomposed

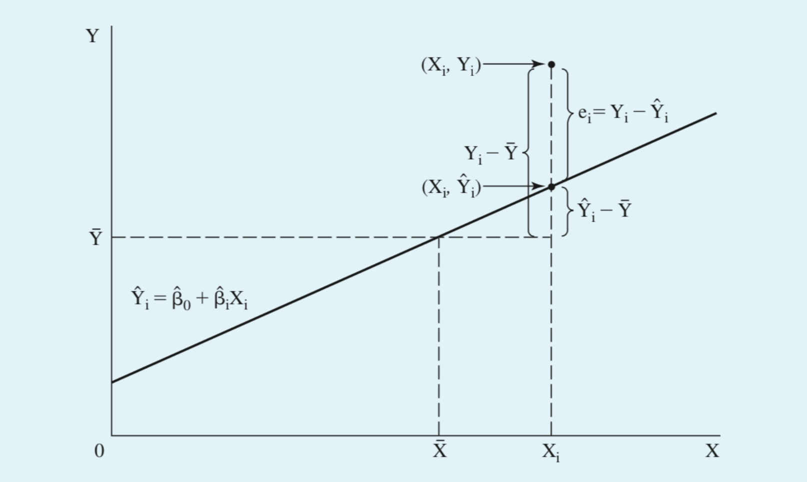
Total sum of squares has two components – variation that can be explained by the regression and variation that cannot be explained.

Decomposition of Variation in Y

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (\hat{Y}_i - \bar{Y})^2 + \sum_i e_i^2$$

Total Sum of Squares (TSS)	=	Explained Sum of Squares (ESS)	+	Residual Sum of Squares (RSS)

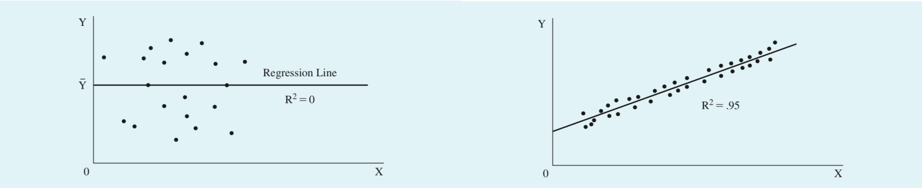
TSS, ESS, RSS and the Regression Line



Coefficient of Determination R^2

To judge the goodness of fit of our model we use R^2

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}$$



Adjusted $\bar{R}^2$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}$$

Problems with R^2

- Adding a variable cannot change TSS, but in most cases the added variable will reduce RSS, so R^2 will rise.
- It also lessens the degrees of freedom (N - K - 1). Fewer degrees of freedom erode the ability to test the model.

Adjusted $\bar{R}^2$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}$$

Problems with R^2

- Adding a variable cannot change TSS, but in most cases the added variable will reduce RSS, so R^2 will rise.
- It also lessens the degrees of freedom (N - K - 1). Fewer degrees of freedom erode the ability to test the model.

Adjusted $\bar{R}^2$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}$$

Problems with R^2

- Adding a variable cannot change TSS, but in most cases the added variable will reduce RSS, so R^2 will rise.
- It also lessens the degrees of freedom (N - K - 1). Fewer degrees of freedom erode the ability to test the model.

Adjusted $\bar{R}^2$

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}$$

Problems with R^2

- Adding a variable cannot change TSS, but in most cases the added variable will reduce RSS, so R^2 will rise.
- It also lessens the degrees of freedom (N - K - 1). Fewer degrees of freedom erode the ability to test the model.

$$\bar{R}^2 = 1 - \frac{\sum e_i^2 / (N - K - 1)}{\sum (Y_i - \bar{Y})^2 / (N - 1)}$$